DOCUMENT RESUME

ED 278 224 FL 016 356

AUTHOR Carroll, Brendan J.
TITLE The Implications for Testing of the Communicative Approach to Language Teaching.
PUB DATE 83
NOTE 11p.; In: Saltarelli, Mario, Ed. The Foreign Language Syllabus--Grammar, Notions and Functions. Urbana, Language Learning Laboratory, University of Illinois at Urbana-Champaign, 1983; see FL 016 352.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Journal Articles (080)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Communicative Competence (Languages); Correlation; *Curriculum Design; Educational Research; *Educational Strategies; Foreign Countries; *Language Tests; Language Usage; Learning Motivation; Qualitative Research; Research Needs; Second Language Instruction; *Test Construction
IDENTIFIERS *Communicative Approach

ABSTRACT

While much has been done in the application of communicative principles to second language instruction, their application to language testing has been less fruitful. Experience with language testing projects in England and other countries suggests that communicative language test design should be based on the following issues and considerations: the distinction and the connections between language and communication; the distinction between language usage and language use; the dangers inherent in too much concern for statistics, especially norm-referenced and correlational techniques; the interdependence of programs, purposes, and testing; the need for a complex, purposive, and strategic model to replace the simplistic four-skills model; the need to accommodate learner motivation and task perception to any external specification of needs; the integration of subjective and objective elements; the assessment of group as well as individual performance; the distinction between test tasks and enabling skills underlying test performance; and the need for a wide range of descriptors derived from a variety of performance criteria. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

THE IMPLICATIONS FOR TESTING OF THE COMMUNICATIVE APPROACH TO LANGUAGE TEACHING

ED278224

BRENDAN J. CARROLL
The Pergamon Press
Oxford, England

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
B. Mainous
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Studies in Language Learning (1982), 4, 46-55.

BEST COPY AVAILABLE

FL016356

0. INTRODUCTION

The current emphasis on the communicative aspects of language has made a considerable impact on language teaching methods and materials, but so far the search for applications of communicative principles to language testing has proved much less rewarding. In this paper, I wish to describe such progress as I and my associates have made in this direction in a number of language testing projects in various parts of the world--the Far East, Mid-East, Africa and Europe--and in particular to two projects I have been involved in at home: the new English Language Testing Service recently brought into use in Britian and overseas,[1,2] and the new Royal Society of Arts Examination in the Communicative Use of English due to be made available in 1981.[3,4]

Three major features of the above projects have been constantly borne in mind during the processes of test development. First, the focus has been on language *use* as well as (and sometimes instead of) language *usage* in specifying the tasks presented in the tests. Although a mastery of linguistic tools (of usage) is certainly a significant factor in assessing communicative performance (in use), and linguistically rated performance doubtless correlates with communicatively rated performance, the two terms are not synonymous. Therefore, if we were to devise effective testing instruments, we would have to incorporate into our test model a systematic and detailed specification of user requirements at the purposive, contextual and strategic levels.[5]

Our second focus arises from the disappointing, even stultifying results of norm-references, especially correlational, statistics whereby highly abstract data-derived schemata have been reified to the point of pedagogic incredibility, so that it has come to be believed that highly complex communicative-linguistic behavior can be adequately assessed via the artificial task of filling gaps in written and spoken texts. The results of one project leads me to believe that elaborate test procedures and data analyses have not been able to solve our test problems because of the indeterminacy of the original, conventional, design model. In short, it has not been possible accurately to decide what our statistical results mean because we had not beforehand adequately described the types of behavior we wished to test.[5]

Our third consideration has been to keep constantly in mind the interrelationship between needs (or purposes), learning programs and test instruments. To operate any single component without regard to its effects on the others would invite the kind of damage which tests and examinations have so often done to education. I believe the value of a test is to be seen not only in its measurement characteristics, such as its reliability and validity, but also in the nature of its contribution to the curriculum system in total.

In truth, the disciplines of grammar and statistics have not so far been used in such a way as to provide satisfactory answers to our language assessment problems; the first discipline being restricted to only a sector of the communication field, the other brought forward to explain linguistic phenomena from data whose bases have not been made sufficiently explicit. In searching to break this deadlock, we have replaced the earlier crude "four-skill" model (see Fig. 2) by a much more complex and purposive scheme derived from that of Mumby[6] and illustrated in Fig. 1.

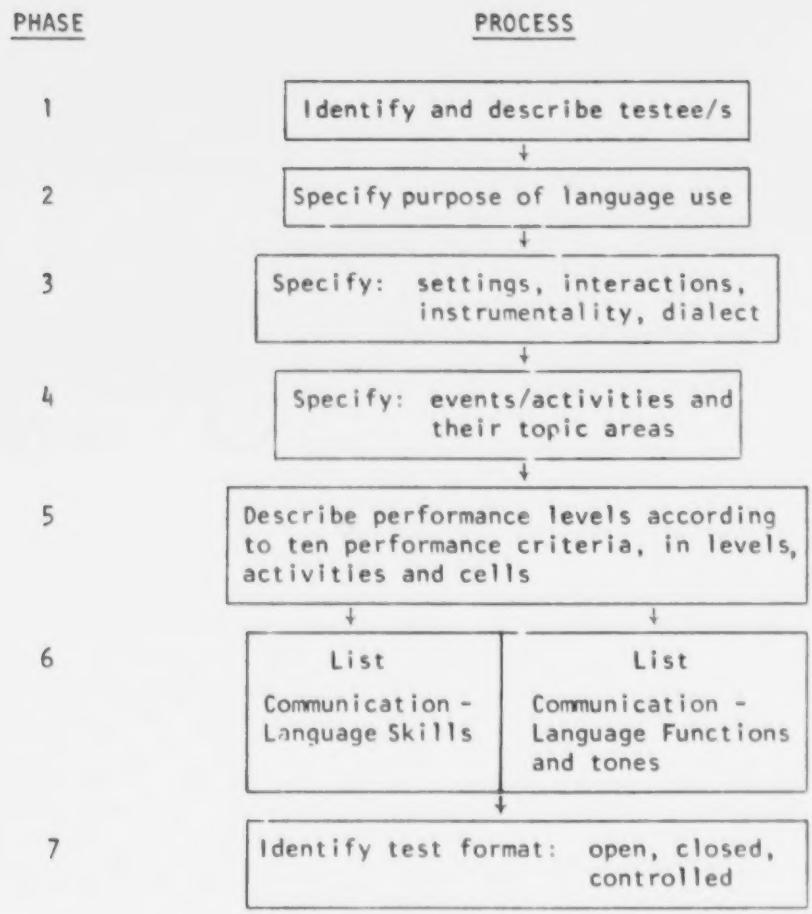

Figure 1. Model for Communicative Tests

Language Aspect / Skill	Pronunciation or Orthography	Grammar	Vocabulary	Other
Listening				
Speaking				
Reading				
Writing				

Figure 2. Commonly-Used Test Model

The procedures for using the communicative model in test development are described more fully in "Testing Communicative Performance."[7] In what follows I discuss certain key aspects of communicative test construction.

1. A SYSTEMS APPROACH

An integrated system of curriculum development will incorporate the specification of purpose, the devising of teacher/learner programs and the use of test instruments to assess how far the learner is achieving his goals. Testing is thus not a self-contained activity, but one affected by purposes and programs and will have its own impact on them. However, in reconciling the various and shifting objectives of the learner, his teacher and the external specifier and tester, I believe we must accept the prior status of purpose (or goal, or need) along which the whole system must be aligned.

2. TEST DESIGN

It is a great temptation to lift from the shelf a ready-made conventional test of "Language Proficiency." Adequate, if at all, for only the crudest of testing purposes, such an instrument is likely to lack relevance and validity insofar as it fails to incorporate critical features of the academic, occupational or social requirements of the learner or of the learning programs devised to meet those requirements. True, it is troublesome to devise afresh special tests for each specific purpose, and the creative leap from needs specification to a workable test instrument is not an easy one, yet no off-the-peg solution, such as the classic cloze test, is able to meet all the communicative objectives revealed by a detailed initial specification.

3. SPECIFIC PURPOSES AND THE LEARNER

To the learner--the principal character in our play--all language needs are to an extent specific; not only because the future settings for his language use are peculiar to him alone, but also because he has a unique personal perception of his own goals and of the learning strategies by which he can best achieve them. Again, a reconciliation is needed; this time between the learner's aspirations, the special constraints of teaching and testing and the often harsh demands for life--resources *are* scarce and jobs and educational courses *do* demand certain levels of performance.

4. AUTHENTICITY

Authenticity has become a controversial term in language teaching not least because it has been thought that a satisfactory program (or test) can be constructed by throwing together a number of tasks which have been observed to be needed on the job. Divorced, however, from their daily settings and jostling together in their new, random environment, these tasks necessarily lose their pristine authenticity and no random set of tasks, however "authentic" in or by themselves, will provide us with the answer to our test design problem. What we need is both to select and to structure the test tasks within

the framework of a purposive, detailed specification model and within a test setting credible to the testee. To achieve these ends, simulation and role-playing are particularly appropriate and make for test operations which, removed from real life to some degree, do have psychological authenticity once the implications of the role "testee" have been accepted. In all cases, the authentic test tasks themselves can only be a small sample of the total task universe, and are to be used to uncover in action the underlying skills and strategies which influence communicative performance.

5. MEASUREMENT

Nowhere more than in language testing can one see the unfortunate results of over-reliance on one or two contributory disciplines. On the one hand, the linguist's delight--the devising of tests to elicit information about acoustic, lexical or syntactic bits treated in an utterly context-free and purposeless fashion. On the other, the tyranny of the statistician, whose lack of insight into communicative phenomena may prove no barrier to his erecting ever more complex psychometric edifices so that, like the eager tourist unable to comment on his travels until the photos have been developed, the ordinary teacher fears to assess his own tests until the statistician has printed out a test data analysis for him. The emphasis on statistical techniques is re-enacted in the pressure for test reliability and the need for variance (or "spread") in test scores. Although a search for variance in itself may be appropriate in cases where one has merely to rank testees, its effects become unhealthy when items acceptable in their content on design grounds are squeezed out because they do not meet criteria of testee discrimination. Thus, the effects of over-emphasis on statistics can be unhealthy, first, by covering up for basic specification faults in the original design model and then by distorting test content by eliminating items just because they do not meet internal discriminatory criteria.

The zenith of dependence on variance is seen in the use of correlation coefficients and factor analysis of correlation matrices. If performance on a particular task, say oral communication, is seen to correlate significantly with performance on another, say gap-filling in written texts, can one of the tasks safely be eliminated from the future test battery? There are, I believe, three main grounds why it would be inadvisable to eliminate a particular test on correlational grounds alone, namely, considerations based on the magnitude of the correlation, the nature of the overlapping variance and the effect on the curriculum system as a whole.

First, as regards the size of the correlation; a coefficient of +0.9 has a predictive value of only 56% better than chance, and one of +0.95 is only 69% better than chance. The commonly-quoted inter-test correlations of the order of +0.6 to +0.8[8] indicate a forecasting efficiency of no greater than 40%; not the predictive value which a wise man would wager too much money on, and hardly sufficient grounds for critical decisions about human beings. Secondly, the nature of the co-occurring variance is also highly suspect. Convergent-discriminant studies[8,9,10] generally indicate excessive amounts of method variance, the variance attributable to testing procedures being about the same as, or even larger than, that attributed to the traits under

scrutiny. A good proportion of any inter-test correlation can thus be discounted as arising from artifacts of the test situation rather than from trait similarity. And this is not even including the broader "error" component. Above all, the test episode is not an isolated one, but is closely related to the whole framework for learning, so that even if the inter-test correlations were error-free, were free from method variance and approached unity (which is not the case) I would still consider it pedagogically irresponsible to eliminate from the test system on correlational grounds a test component which had genuine a priori design grounds for inclusion.

The adducing of a "General Verbal Factor," revealed by factor analysis, is even more suspect ground for homogenizing a test battery, as the alleged general factor can quite easily arise from the indeterminacy of the original model, the analysis not so much proving the existence of a general factor as being unable to distinguish the true factor pattern because of faulty design and/or test realizations. To return to the main point, then, I believe we have not been accurately measuring linguistic behavior because we have lacked an adequate purposive, detailed specificatory model. When we have not been testing properly, we have not been adequately describing the behavior we have aspired to measure.

6. BAND DESCRIPTIONS

One instrument we have used to maintain a degree of measurement with a meaningful description of performance has been the time-honored language scale, or band, system. The elements of the band system are levels, activities, cells and descriptors and the band levels can be brought into scale with objectively scored tests to provide a meaningful profile of performance for any testee. The band elements are described below.[11]

Levels: the number of scale intervals may be from 3 (e.g., good, average, poor) to anything up to 11 (e.g., the Foreign Service Institute Scale with categories 1 to 5 including subdivisions from 1 to 4; and possibly 0 and 0+). Our own projects draw on a 9 point scale (9 down to 1).

Activities: generally a mix of two or more modes such as conversation (listen - speak), note-making (read - write), etc., rather than just one of the traditional "four skills."

Cells: at the intersection of each level and activity whose performance descriptions are inserted.

Descriptors: a spelling-out of performance at an appropriate level of delicacy according to a set of performance criteria from such an inventory as: size, complexity, range, speed, flexibility, accuracy, appropriacy, independence, repetition and hesitation.[6]

We can thus provide either a brief outline of types of performance, as in the FSI Language Proficiency ratings, or a very detailed description as in the Oral Interaction scale described in "Testing Communicative Performance."[7] By

a simple scaling procedure, the band performance can be linked with closed-ended, scored tests so that any testee can be given an individual profile, as shown in these four examples of English Language Testing Service profiles:

Sample ELTS Profiles

Test / Candidate	G 1	G 2	M 1	M 2	M 3	Overall
4729	6.0	5.5	7.5	6.5	6.0	6.5
4730	3.0	2.5	3.0	2.0	2.0	2.5
4731	6.5	8.0	8.5	7.0	8.0	7.5
4732	5.5	5.5	6.0	5.5	5.0	5.5
	(all scores to nearest half-band)					

A candidate's profile is matched with minimum "job profiles" for various types of training courses to decide either his suitability for a course or his language tuition requirements. Thus the subjective estimates of performance are reinforced by objective test scores, combining the most useful features of the two types of measure.

7. GROUP ACTIVITIES

Much of the discussion on testing concerns *individual* performance yet, communication being primarily an inter-personal phenomenon, there is much to be said for assessing communicative behavior on a pair or group basis. We are devising scales for judging the effectiveness of intra-group communication in accomplishing given tasks. In this exercise if an individual were to be judged, it would be according as he fulfills his role, however humble, in assisting the group to achieve its objectives. The inclusion of such group tasks in the examination framework is an important step in motivating group work in language learning programs; indeed the curriculum effects may well be more significant than the measurement element itself.[12]

8. LANGUAGE AND COMMUNICATION

To repeat the crucial point made in our introduction, tests which require the manipulation of language forms can tell us only half of the story because they fail to include major parameters of linguistic behavior--the strategies, purposes and the rules of language use in defined settings. It is profitable, therefore, to establish beforehand whether, or to what extent, a test should contain real (or at least realistic, simulated) tasks or whether the focus

should be on the mastery of those skills which enable the tasks to be carried out. Even then, such "enabling" skills are best tested in the context of realistic test tasks.

9. JUDGING COMMUNICATIVE VALUE

Whereas linguistic accuracy can be judged relatively simply, communicative value depends on a range of factors many of which will be not basically linguistic. To assess oral performance, for example, we may vary the difficulty of the task and/or judge the value of the performance according to such variables as:

Channel: face-to-face, phone, tape, film.

Fluency: speed, coherence, absence of hesitation or of undue repetition.

Dialect: the language variety used.

Affect: The appropriacy and facility of conveying or comprehending attitudes.

Text type: register, style, format of language.

Status: the relative dominance and social distance of the participants.

Topic expectancy: familiarity with the topic at hand.

"Noise": both acoustic *re* distortion of sound, and psychological *re* gaps, interruptions and interference.

Length: duration of the total episode and of the relative contributions of the participants.

Non-language support: use of visuals, diagrams, symbols, media, body movement and understanding of situations.

Diversity: number of topics, focuses and switches.

Linguistic accuracy: acceptability/correctness of usage, pronunciation, vocabulary.

Overall effectiveness: adequacy in handling the interaction as a whole.[13]

We are developing studies to discover how variables such as the above contribute to effectiveness in oral interaction and how they can be related and weighted in assessment, and thus used in both the devising of oral tests and the assessment of oral performance. Such research, I believe would do much more to improve our insights into communicative behavior than ever more studies into the non-communicative discrete categories of listening, speaking, reading and writing judged in terms of the linguistic characteristics of pronunciation, vocabulary, grammar and fluency.

10. CONCLUSION

The main points of my thesis may be condensed as follows:

- The distinctions, as well as the relationships between communication and language.
- The distinction between usage and use.
- The dangers of over-riding concern with statistics, especially norm-referenced and correlational techniques.
- The crucial inter-dependence of programs, purposes and testing.
- The need for a complex, strategic and purposive model to replace the currently-used simplistic "four-skill" model.
- The need to accommodate learner motivation and task perception to any externally-devised specification of needs.
- The development of band theory and the integration of subjective and objective elements in a test.
- The desirability of developing criteria to assess the effectiveness of *group*, as well as individual, performance.
- The distinction between test tasks and the enabling skills which lie behind test performance.
- The need for a wide range of descriptors derived from a variety of performance criteria.

To conclude with an analogy. Many of the critical points just outlined can be discussed in terms of the driving test. It may be found, for example, that there is a fairly high correlation between a learner's actual driving performance on the road and his performance in a simulated car interior where he reacts to filmed stimuli. Would we be justified in doing away with an actual driving test on the basis of this correlation? Obviously, for early training and the assessment of learning potential, such a simulated setting would have many advantages in that the learner is able to develop the enabling skills of driving in fairly stress-free conditions. But this is a very different matter from leaving all the assessment at the level of enabling skills. The basic constraints of a test situation are also illustrated in the driving test, which cannot completely replicate actual driving conditions. First, there is an observer on hand; then, night driving and motor-way driving are not allowed; again, problems encountered in driving very long distances over an extended period can rarely be revealed in a short test. As a measurement procedure, perhaps too much depends on the subjective impressions and the personal prejudices of the driving inspector, and it is unduly affected by nervousness induced by the test situation leading to a typical over-reaction to small incidents.

Such are the points touched on in my discussion on language testing; and clearly there is no easy optimum formula to balance out such a mass of considerations. Perhaps much will have been achieved if we realize the primitive stage of language testing at present and the magnitude of the work still needed for us to understand and, thus, to assess, linguistic-communicative behavior.

REFERENCES

[1]Carroll, B.J., 1978, *Specifications for an English Language Testing Service.* London: British Council.

[2]*Handbook.* London: English Language Testing Service, British Council (in preparation).

[3]Morrow, K.E., 1978, *Techniques of Evaluation for a Notional Syllabus.* London: Royal Society of Arts.

[4]Royal Society of Arts, 1980, *Specifications: Examinations in the Communicative Use of English as a Foreign Language.* London: Royal Society of Arts.

[5]Carroll, B.J., 1980, "The Relative Contributions in Testing of Linguistic and Communicative Factors," paper delivered to Colorado TESOL.

[6]Munby, J.L., 1978, *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

[7]Carroll, B.J., 1980, *Testing Communicative Performance.* Oxford: Pergamon Press.

[8]Clark, J.L.D., ed., 1978, *Direct Testing of Speaking Proficiency.* Princeton, ETS, 193 ff.

[9]Campbell, D.I., and D.W. Fiske, 1967, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," in Mehrens, W.A., ed., 1967, *Principles of Educational and Psychological Measurement.* Chicago: Rand McNally.

[10]Carroll, B.J., 1979, "A Communicative View of Convergent-Discriminant Validation," Typescript Paper to TESOL Current Study.

[11]Carroll, B.J., 1979, "A Common Standard Scale for Language Assessment," paper prepared for international conference at ETS, Princeton.

[12]Note: References to "Crescent" English Language programme, *and* Oxford Modern Language Project.

[13]Note: List of variables has been devised in discussion with J.L.D. Clark and P.E. Woodford of ETS Princeton on common language scale project referred to in (11) above.